DOCUMENT RESUME

ED 201 660 TM 810 253

AUTHOR Strasler, Gregg M.
TITLE A Case in Point: A Discrimination Index for Criterion-Referenced Tests.
PUB DATE Apr 80
NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 8-10, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Discriminant Analysis; Higher Education; *Mastery Learning; *Norm Referenced Tests; Scores; Statistical Significance; *Test Theory
IDENTIFIERS *Discrimination Index; Point Biserial Correlation; Tetrachoric Correlation; Variability

ABSTRACT

The relationship between classical discrimination indices (CDI) and criterion-referenced discrimination indices (CRDI) and the appropriateness of each for use on criterion-referenced tests are investigated. A CRDI is proposed that attempts to separate those who master material from those who do not master material. A 26 item multiple-choice formative test (as well as a parallel true-false test) was given to 205 college undergraduates. The students were also given one period of instruction, and a table of specifications was made for their unit of learning. There were 32% "masters" (achieving 80% correct) and 68% "non-masters" within the sample on the multiple-choice test. The CDI and the CRDI were computed and compared with point biserial and tetrachoric inter-item correlations respectively. All 26 items were significant at the 0.05 level with respect to the CDI, whereas 22 items were significant with respect to the CRDI. There were 22 items in common that were significant for both the CDI and CRDI. It appears that classical indices tend to include more items than criterion-referenced indices. Futhermore, criterion-referenced items tend to be a subset of norm-referenced items with few exceptions. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED201660

A Case in Point: A Discrimination Index for Criterion-Referenced Tests

Gregg M. Strasler
S. C. Region V Educational Services Center
and
University of South Carolina

Abstract

The purpose of this study was to investigate the relationship between classical discrimination indices and criterion-referenced discrimination indices and the appropriateness of each for use on criterion-referenced tests. A discrimination index was proposed that separated "masters" from "non-masters" in contrast to the classical approach. A criterion-referenced formative test was administered to 205 college undergraduates at a state supported university and the data were analyzed in a variety of ways. Classical indices tended to include more items than criterion-referenced indices. Furthermore, criterion-referenced items tend to be a subset of norm-referenced items with few exceptions.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. M. Strasler

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 253

A paper presented at the annual meeting of the National Council on Measurement in Education, Boston, 1980.

A Case in Point: A Discrimination Index for Criterion-Referenced Tests[1]

Gregg M. Strasler
S. C. Region V Educational Services Center
and
University of South Carolina

Objectives/Theoretical Framework

The purpose of this study was to investigate the relationship between classical discrimination indices and criterion-referenced discrimination indices and the appropriateness of each for use on criterion-referenced tests.

Within the last decade, the interest in mastery learning strategies and competency-based education has led to an increased attention to criterion-referenced tests. With the increased importance of criterion-referenced tests, a controversy has arisen as to the feasibility of using conventional item analysis in isolating those items that may require further revision. It has been argued that item variance and test variance are not necessary characteristics for criterion-referenced tests, although they are necessary for norm-referenced tests. Popham and Husek (1969) have argued that the test items for criterion-referenced tests and the tests themselves may have no variance and therefore traditional methods of empirical evaluation of test items and tests are invalid for criterion-referenced tests. Popham and Husek (1969) conclude, "With criterion-referenced tests, variability is irrelevant . . . Variability is not a necessary condition for a good criterion-referenced test" (p.3).

Woodson (1974) maintains that the position held by Popham and Husek is in error because it considers sample statistics as the criteria for evaluating items and tests. Woodson states that within a particular sample, an item or test may have no variance, but in the population of observations for which the test was designed, calibrated, and evaluated, both items and tests must have variance. Woodson (1974) concludes, ". . . item variance and discrimination are essential. In short, (1) items and tests must be evaluated for the range of the characteristic for which they will be used, and (2) items and tests which give no variability in this population of observations, give no information and are therefore not useful" (p.64).

In response to Woodson (1974), Millman and Popham (1974) contend that the population of observations originally prescribed for criterion-referenced tests is a domain of items and the responses of a single individual to them. Millman and Popham (1974) state, ". . . item or score variability is an unnecessary characteristic of criterion-referenced tests as they have been traditionally conceived, namely, as measures of well-defined classes of examinee behaviors" (p.137). Woodson (1974) states that Millman and Popham (1974) tend to see the outcome of criterion-referenced measurement as an all or nothing decision (the classifying of an individual's knowledge as nonmastery or mastery). According to Woodson (1974), ". . . individuals differ on any characteristic of interest, including all dimensions of learning and achievement" (p.140).

---

[1]This paper was originally conceptualized in 1974 by the present author and Dr. Lorin W. Anderson of the University of South Carolina.

Both sides of the controversy readily acknowledge that score variability will almost always be found in practice (cf. Ebel, 1977; Haladyna, 1974, 1976). As stated by Linn (1979), "The crux of the difference lies in the importance attached to the notion that variability is not a necessary consequence of a good competency measure . . . In part the debate may be viewed as one of interest 'in theory' if not in practice" (p.92). Although the issue of score variability remains largely unresolved, it has been important not only to theoretical discussions but to the search for alternate procedures involving indices of reliability, validity, and discrimination.

In dealing with the issue of score variability in criterion-referenced testing, it should also be noted that there is a difference between an "item" and a "construct" in testing. One test item may not accurately measure the construct it was intended to measure. For example, in a graduate course at the University of South Carolina, an attempt was made to ascertain the amount of information provided by asking two (2) parallel questions for every "construct" being measured. Therefore, only when both parallel items were answered correctly (RR) or incorrectly (WW) by a student could accurate information be obtained on a particular construct. It was estimated that approximately seventy-three percent (73%) of the item pairs provided "accurate" information (see Table 1).

In conclusion, some type of discrimination index is needed for criterion-referenced tests (cf. Ebel, 1977; Pettie & Oosterhof, 1976; Scannell & Tracy, 1975). A discrimination index is proposed that attempts to separate those who mastered the material from those who did not master the material in contrast to the classical approach.

Methods/Techniques

A criterion-referenced formative test (see Table 2) was constructed according to the principles of Bloom, Hastings, and Madaus (1971). The sample (205 college undergraduates) was tested on Chapter 10 ("Of the Smell of Nell: Infancy") in Guy R. Lefrancois' book, Of Children. In this chapter, the development of the infant's intellect is examined. Before the twenty-six (26) item multiple-choice formative test (as well as a parallel true-false test) was administered, the sample was given a rationale for formative testing and mastery learning. They were also given one period of instruction by their instructor on Chapter 10. A table of specifications was made for this unit of learning. Item content for levels of knowledge and comprehension was checked and verified by one independent judge.

There were approximately thirty-two percent (32%) "masters" as opposed to sixty-eight percent (68%) "non-masters" within the sample on the multiple-choice formative test. A criterion score of twenty-one (21) items correct (approximately 80% correct) was set for mastery.

The data were analyzed in a variety of ways. The classical discrimination index was computed by subtracting the proportion of the bottom twenty-seven percent (27%) of the sample getting an item correct from the proportion of the top twenty-seven percent of the sample getting an item correct. A point biserial correlation of each item with the total test score was investigated.

The criterion-referenced discrimination index was computed by subtracting the proportion of "non-masters" of the sample getting an item correct from the proportion of "masters" of the sample getting an item correct. A standard error of measurement was computed (SEM = 2.04) and only those "masters" who scored one standard error of measurement or more above the criterion score of eighty percent (80%) correct were compared to only those "non-masters" who scored one standard error of measurement below the criterion score of eighty percent (80%) correct.[1] A tetrachoric inter-item correlation of each item with "mastery" versus "non-mastery" was investigated.

Results/Conclusions

Ebel (1965) has suggested that an item should have an index of discrimination of a minimum of 0.30 before it is included in a test. Using this criterion more items were included in the multiple-choice test using a classical or norm-referenced discrimination index (see Table 3). Of the twenty-six (26) items, fifteen (15) were included using a norm-referenced discrimination index whereas twelve (12) were included using a criterion-referenced discrimination index. These items account for approximately fifty-eight percent (58%) and forty-six (46%) of the total items using norm-referenced and criterion-referenced discrimination indices respectively. Also, it should be noted that there were no items which would be included by the criterion-referenced discrimination index that would not be included by the classical or norm-referenced discrimination index. In contrast, there were three (3) items included by the classical discrimination index that would not be included by the criterion-referenced discrimination index. There were twelve (12) items in common that fit both the classical and criterion-referenced discrimination indices.

When classical and criterion-referenced indices were compared with point biserial and tetrachoric inter-item correlations respectively, similar results were found. All twenty-six (26) items were significant at the 0.05 level with respect to classical or norm-referenced indices whereas twenty-two (22) items were significant with respect to criterion-referenced indices. There were twenty-two (22) items in common that were significant for both classical and criterion-referenced indices. It appears that classical or norm-referenced indices tend to include more items than criterion-referenced indices. Furthermore, criterion-referenced items tend to be a subset of norm-referenced items.

---

[1]Nassif (1978) presents some support of this technique in a paper illustrating procedures used to validate and set standards for criterion-referenced teacher certification exams. Nassif (1978, p. 13) states:

In order that examinees not be denied professional teaching licenses because of errors of measurement, the minimum cutoff score determined by expert judges was adjusted by the lower half of the standard error of measurement interval computed at 95% confidence. In this way the Adjusted Cutoff = Original Cutoff - Standard Error of Measurement (SEM = 1.96 S.D. $\sqrt{1 - REL_{\alpha}}$). Applicants performing just below the original cutoff level would now be considered minimally competent, after adjustments for measurement errors. This adjustment of the cutoff score due to measurement error has been supported in South Carolina (South Carolina Education Association et al. v. State of South Carolina et al., Civil Action No. 75-1610, District of South Carolina, 1977).

Educational Importance of the Study

Whether or not classical test theory can be applied to a criterion-referenced framework has been a debated issue in recent years. One camp advocates that the concept of variability in test scores is irrelevant with criterion-referenced tests (e.g., Popham and Husek, 1969; Millman and Popham, 1974). The other camp emphasizes that variability has been observed in criterion-referenced testing and is an important concept to be considered (e.g., Woodson, 1974; Haladyna, 1974). The question remains unresolved as to whether classical test measurement is appropriate for the evaluation of criterion-referenced tests.

Since most existing criterion-referenced tests are, in fact, teacher-constructed, it is important to include some kind of discrimination index. Since items are useful in making inferences concerning the individual's ability vis a vis a particular construct and are not equivalent to the construct itself, faulty items may give under- or over-estimates of the individual's ability on that construct. Given the purpose of criterion-referenced measures a criterion-referenced discrimination index is called for. The criterion-referenced discrimination index is an index which is virtually independent of the sample under consideration. As such it seems to be more precise in selecting appropriate items that do, in fact, separate those possessing the criterion level of ability from those possessing the pre-set level.

Table 1

Percentage of Information Obtained on Parallel Test Items

Multiple Choice/True-False Patterns

Item	RR	WW	Accurate Information	RW	WR	Inaccurate Information
1	64.9	11.2	76.1	21.0	2.9	23.9
2	90.7	2.9	93.6	4.4	2.0	6.4
3	85.4	6.3	91.7	1.5	6.8	8.3
4	8.8	10.7	19.5	79.0	1.5	80.5
5	67.3	15.1	82.4	12.7	4.9	17.6
6	83.4	8.3	91.7	3.4	4.9	8.3
7	86.8	2.4	89.2	10.2	0.5	10.7
8	43.9	27.8	71.7	6.3	22.0	28.3
9	83.9	-	83.9	14.1	2.0	16.1
10	89.2	0.5	89.7	0.5	9.8	10.3
11	67.3	11.7	79.0	14.7	6.3	21.0
12	67.3	9.8	77.1	18.5	4.4	22.9
13	45.4	18.0	63.0	7.8	28.8	36.6
14	12.2	53.7	65.9	27.3	6.8	34.1
15	29.3	35.1	64.4	20.5	15.1	35.6
16	60.0	7.3	67.3	22.4	10.3	32.7
17	69.3	17.1	86.4	5.3	8.3	13.6
18	86.3	3.9	90.2	3.9	5.9	9.8
19	46.3	24.9	71.2	10.2	18.5	28.7
20	35.6	31.3	66.9	2.9	30.2	33.1
21	24.4	35.1	59.5	31.2	9.3	40.5
22	42.0	19.0	61.0	5.8	33.2	39.0
23	46.8	16.6	63.4	12.2	24.4	36.6
24	40.5	34.1	74.6	8.8	16.6	25.4
25	73.7	0.5	74.2	0.9	24.9	25.8
26	3.4	29.8	33.2	64.9	1.9	66.8
Total	55.9	16.7	72.6	15.8	11.6	27.4

Note. Information was identified as "accurate" only when both parallel items (multiple choice/true-false) for a particular "construct" being measured were answered correctly (RR) or incorrectly (WW).

Chapter 10: Of the Smell of Nell: Infancy
Learning Exercise

DIRECTIONS: This exercise consists of two (2) parts containing a total of fifty-two (52) questions. When you have decided upon the correct answer to a question, blacken the corresponding space on the separate answer sheet that is provided. You have one hour to complete this exercise.

PART I: MULTIPLE CHOICE

1. The transmission of messages may be defined as:

   a. thought
   b. language
   c. intelligence
   *d. communication

2. Language codes of lower class children are described by Bernstein as being:

   a. elaborated
   b. formal
   *c. restricted
   d. public

3. The first stage of development in Piaget's classification is termed as:

   a. preoperational
   b. concrete operations
   c. formal operations
   *d. sensorimotor

4. The belief that language not only allows people to express their thinking but is also largely responsible for determining those thoughts is expressed by:

   a. Bernstein
   *b. Whorf
   c. Baratz
   d. Barnett

5. Imitating people or events in their absence is called:

   a. enactive imitation
   b. secondary imitation
   *c. deferred imitation
   d. primary imitation

6. Language codes of middle class children are described by Bernstein as being:

   *a. elaborated
   b. formal
   c. restricted
   d. public

7. Complex arrangements of arbitrary sounds that have accepted referents and can therefore be used for communication among humans are referred to as:

*a. language
b. communication
c. thought
d. intelligence

8. The belief that children represent their world by the activities they perform with it or on it is called:

*a. enactive representation
b. secondary representation
c. deferred representation
d. primary representation

9. The relatively meaningless sounds that young infants repeat are referred to as:

a. stuttering
*b. babbling
c. bilingualism
d. syntax

10. It appears that the disadvantaged child is not necessarily less advanced in language development, but that in many cases he has simply developed a different language:

a. development
*b. code
c. concept
d. communication

11. The roots of language are found in the infant's:

*a. babbling
b. crying
c. gestures
d. meaningful speech sounds

12. The sensorimotor period as described by Piaget lasts from:

*a. birth to two years
b. birth to concrete operations
c. birth to five years
d. one to two years

13. When an object is internalized or becomes part of one's cognitive structure, those activities directly associated with the real object comprise the individual's representation of the object. Bruner terms this:

*a. enactive representation
b. secondary representation
c. deferred representation
d. primary representation

14. Communication, thought, and intelligence have a great deal to do with the:

    a. formality of language
    b. instrumentation of language
    *c. power of language
    d. concept of language

15. The child's eventual achievement of object permanence is inferred from what Piaget labels:

    a. enactive imitation
    b. secondary imitation
    *c. deferred imitation
    d. primary imitation

16. Lenneberg, Nichols, and Rosenberber (1964) observed that when mongoloid children were still crawling, regardless of age, their language development was usually at the stage of:

    a. sensorimotor development
    b. talking
    *c. babbling
    d. concrete operations

17. The prespeech stage ends with the development of:

    a. babbling
    b. crying
    c. gestures
    *d. meaningful speech sounds

18. All the sounds of every language in the world are uttered in the infant's:

    *a. babbling
    b. crying
    c. gestures
    d. meaningful speech sounds

19. In summarizing his findings, Bernstein observed that the language of the lower-class family has certain characteristics. Which one of these characteristics was <u>not</u> observed?

    a. It tends to be liberally sprinkled with idioms
    *b. It tends to be more personal
    c. It is more global and less precise
    d. It is employed to control emotion

20. The first step in the acquisition of language probably involves:

    a. babbling
    b. repetition of sounds
    *c. crying
    d. gestures

21. The Whorfian hypothesis implies that the power of language has a great deal to do with:

    a. communication
    b. communication and intelligence
    *c. thought and intelligence
    d. preverbal thinking

22. Osgood (1957) describes the process of acquiring language as a circular reflex. The first step in this circular reflex involves:

    *a. babbling
    b. meaningful speech sounds
    c. crying
    d. gestures

23. Baratz (1969) found that Negro and Caucasian children performed equally well when confronted with language codes that were outside their primary code; while the level of performance in standard English was significantly different for the two groups, their language development was actually similar. In other words, these findings suggest that:

    a. disadvantaged children fare less well in language tests
    *b. lower-class children are not deficient in language development
    c. lack of facility with the school language code remain deleterious to successful performance in school
    d. Caucasian children have less trouble with nonstandard English codes than Negro children have with standard English codes

24. The acquisition of language through auditory feedback is described by Osgood (1957) which demonstrates when the child says his first word. In other words, the sound "dah" becomes a word when the child:

    a. emits the sound, "dah;" hears it, and repeats it because he hears it
    b. becomes capable of responding to auditory feedback from his own vocalizations
    c. repeats the sound, "dah," many times before it becomes associated with an object
    *d. vocalizes the sound, "dah," which has come to mean something to him

25. In terms of language acquisition, what is the most important similarity between the findings of Lenneberg, Nichols, and Rosenberber (1964) with mongloid children and Piaget's observation that a child uses words relating mainly to the "here and now" in the early stages of speech development?

    *a. There appears to be a positive correlation between motor capabilities and language development
    b. The ability to conceptualize the environment is reflected in the infant's mushrooming language development
    c. The child's understanding of objects corresponds to the activities he performs with them
    d. The greatest mystery of the process of acquiring language is the inability to explain how the child makes such extraordinary progress

26. One hypothesis states that language not only allows people to express their thinking but is also largely responsible for determining those thoughts. In other words,

a. it would be necessary to accept that there is preverbal thinking in children
b. the power of language has more to do with communication than with thought and intelligence
*c. we think those thoughts for which we have language
d. the power of language has more to do with thought and intelligence than with communication

PART II: TRUE-FALSE

27. T *F The transmission of messages may be defined as language.

28. *T F Bernstein describes language codes of lower class children as being restricted.

29. *T F The first stage of development in Piaget's classification is termed as sensorimotor.

30. *T F The Whorfian hypothesis express the belief that language not only allows people to express their thinking but is also largely responsible for determining those thoughts.

31. T *F Imitating people or events in their absence is called secondary imitation.

32. *T F Bernstein describes language codes of middle class children as being elaborated.

33. T *F Complex arrangements of arbitrary sounds that have accepted referents and can therefore be used for communication among humans are referred to as thought.

34. *T F The belief that children represent their world by activities they perform with it or on it is called enactive representation.

35. T *F Stuttering may be defined as the relatively meaningless sounds that young infants repeat.

36. *T F It appears that the disadvantaged child is not necessarily less advanced in language development, but that in many cases he has simply developed a different language code.

37. T *F The roots of language are found in the infant's crying.

38. T *F The sensorimotor period as described by Piaget lasts from birth to concrete operations.

39. *T F Bruner states enactive representation occurs when those activities directly associated with the real object comprise the individual's representation of the object.

40. T *F Communication, thought and intelligence have a great deal to do with the instrumentation of language.

41. T *F Piaget defines secondary imitation as the child's eventual achievement of object permanence.

42. *T F Lenneberg, Nichols, and Rosenberber (1964) observed that when mongloid children had begun to walk, they had also begun to talk.

43. *T F The prespeech stage ends with the development of meaningful speech sounds.

44. *T F All the sounds of every language in the world are uttered in the infant's babbling.

45. *T F In summarizing his findings, Bernstein observed that the language of the lower-class family is employed to control emotion.

46. *T F The first stage in the acquisition of language probably involves crying.

47. T *F The Whorfian hypothesis implies that the power of language has a great deal to do with preverbal thinking.

48. *T F Osgood (1957) states babbling is the first step in the process (circular reflex) of acquiring language.

49. T *F According to Baratz (1969), Caucasian children have less trouble with nonstandard English codes than Negro children have with standard English codes.

50. T *F According to Osgood (1957), the sound, "dah," becomes a word when the child emits the sound, "dah," hears it, and repeats it because he hears it.

51. *T F There appears to be a positive correlation between motor capabilities and language development.

52. *T F One hypothesis states that language not only allows the transmission of thought, but that it is also largely responsible for determining the substance of that thought.

Table 3

Discrimination and Correlation Indices for Multiple Choice Test

	Discrimination[1]		Correlation[2]	
Item	Norm-referenced	Criterion-referenced	Norm-referenced	Criterion-referenced
1	.26	.20	.24*	.19*
2	.13	.09	.23*	.09
3	.17	.15	.22*	.11
4	.22	.17	.26*	.21*
5	.46*	.33*	.44*	.36*
6	.35*	.24	.38*	.26*
7	.07	.04	.15*	.03
8	.56*	.54*	.45*	.51*
9	.04	.03	.14*	.04
10	.24	.16	.36*	.14*
11	.36*	.26	.36*	.23*
12	.22	.22	.26*	.19*
13	.62*	.58*	.52*	.56*
14	.49*	.57*	.39*	.45*
15	.40*	.61*	.37*	.46*
16	.33*	.25	.36*	.17*
17	.15	.28	.16*	.29*
18	.18	.15	.28*	.16*
19	.45*	.44*	.35*	.46*
20	.33*	.39*	.31*	.30*
21	.45*	.35*	.34*	.30*
22	.42*	.35*	.30*	.36*
23	.33*	.32*	.26*	.23*
24	.44*	.45*	.31*	.47*
25	.22	.11	.21*	.17*
26	.60*	.49*	.51*	.49*

[1] * = index of discrimination of .30 or greater

[2] * = r significant at .05 level

References

Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Ebel, R. L. A response to attacks on testing. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, 1977.

Ebel, R. L. Measuring educational achievement. Englewood Cliffs, N. J.: Prentice-Hall, 1965.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99.

Haladyna, T. The paradox of criterion-referenced measurement. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1976.

Lefrancois, G. R. Of children. Belmont, California: Wadsworth, 1973.

Linn, R. L. Issues of reliability in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based education. Washington, D. C.: NCME Monograph, 1979.

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.

Nassif, P. M. Standard-setting for criterion-referenced teacher licensing tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Toronto, 1978.

Pettie, A. A., & Oosterhof, A. C. Indices of item adequacy for individually administered mastery tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Scannell, D. P., & Tracy, D. B. Testing and measurement in the classroom. Boston: Houghton Mifflin, 1975.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests: A reply. Journal of Educational Measurement, 1974, 11, 139-140.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64.